This paper is an overview of the state-of-the-art of methods for the Chinese word segmentation task, in particular some investigations of overlapping ambiguity distribution in the corpus, and the overlapping ambiguity detection coverage of the FMM+BMM method.

Segmentation Problems in Chinese Processing
Changning Huang

语言文字应用 1997 年第 1 期(总第 21 期)创刊五周年纪念号

中文信息处理中的分词问题*

黄昌宁

提要 在中文信息处理的众多应用领域中,从最底层的键盘、语音和字符识别等各类汉字输入方法,到最高层的各种汉语理解系统,都不可能完全摆脱汉语文本分词处理的困扰。分词问题已成为当前中文信息处理的一个瓶颈。没有一种公认的分词标准,是人和计算机共同面临的困难。如果在这个问题上不能尽快达成共识,那么在词表和带标注的语料库等重要信息资源上就不可能做到共享与复用,势必造成重复开发的严重浪费。当务之急是制定一份与分词规范相配合的汉语通用词表。

谈及汉语文本的分词问题,不由得想起 1990 年 9 月香港《语文建设通讯》第 30 期上李友仁先生对《信息处理用现代汉语分词规范(草案)》所发表的意见。他说:"在计算机上对现代汉语作信息处理,无非是在计算机上输入输出方块汉字。"又说:"输入现代汉语的汉字文稿是字和字串与标点符号连续输入的,不是分词输入的。输出的现代汉语汉字文件也是字和字串与标点符号连续排版的,不是分词排版的。……既然如此,'分词'还有什么实际意义呢?'分词规范'还有什么实际意义呢?如果都无实际意义,'切分标准、原则'就更无实际意义了。"

对李先生的见解,有必要澄清如下两个概念:第一,汉语信息处理决不仅限于在计算机上输入输出汉字。汉语信息处理又称中文信息处理,是指"用计算机对汉语的音、形、义等信息进行处理",包括"对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作与加工"。[1]解决计算机的汉字输入输出问题曾经是早期中文信息处理界的一项具有历史意义的重要任务,这个时期因此得名为"字处理阶段"。至 80 年代中,国家有关部门陆续制定和颁布了信息处理用汉字编码字符集、汉字点阵字模集,一批实用的汉字编码(键盘)输入系统也先后问世。汉字进入计算机的梦想得以实现,标志着中文信息处理已从"字处理阶段"迈入"词处理阶段"。第二,由于汉语中同音字太多,有的音节的同音字高达 100 多个,基于拼音的汉字编码方法必须依靠"以词定字"甚至"以句定词"的手段来区分同音字和同音词。到 80 年代末期,不具备词处理技术的汉字编码输入方法几乎已绝迹。可见,即使仅仅为了解决计算机的汉字输入问题,也不能说汉语文本的分词是没有实际意义的。

一 分词问题的重要性

文稿中词与词之间没有明确的分隔标记是汉语和大多数东方语言的一个显著特点。但词是"最小的能独立运用的语言单位",[1]要实现中文信息处理的各项任务,分词问题的重要性就显得异常突出了。下面让我们从不同角度考察一下分词对中文信息处理的作用。

* 国家自然科学基金重点项目资助。

1. **同音字的据音辨字**　据统计,普通话共有414个不分声调的音节,带上4个声调的音节也只有1295个。以7000个汉字计算,平均每个带调音节对应5.4个汉字,这就是汉字中出现大量同音字的缘故。例如,不分声调的音节yi有131个同音字,ji有121个,yu有115个,xi有102个,等等。[2]在书面语中,大多数同音字都可依靠"以词定字"的手段来加以区别,如"示、世、市、式、事、适"等都读shì,但当它们分别出现在"表示、世界、城市、方式、事情、合适"等词中,[2]就不难区分了。即使是由同音字形成的单字词(如"是"),也可根据前后词的语境来加以辨识。同音字辨识是汉字编码拼音输入方法和汉语语音识别等应用中的重要技术。

2. **多音字的据字辨音**　多音字指一个汉字有几种不同读音,从而形成一字多音的现象。如·好(hǎo)人|爱好(hào);行(xíng)人|人行(háng)(人民银行);率(shuài)领|效率(lǜ);学校(xiào)|校(jiào)对,等等。同理,汉语中的多音字大多数也要靠"以词定音"或句中前后词语境来实现字音的辨识。汉语多音字的字音辨识是汉语文本-语音转换课题的关键技术之一,其实现离不开文本的自动分词技术。

3. **汉字的简-繁体转换**　从简体字到繁体字的转换所遇到的困难也是一对多,即一个简体字有可能对应多个繁体字,一样要靠"以词定字"或上下文来解决。例如:简体字"淀"可对应于传承字"淀"和繁体字"　",如白洋淀|　粉;简体字"干"可对应于传承字"干"和繁体字"乾",如干涉|干戈|乾燥|乾脆|乾娘;简体字"发"可对应于繁体字"发"和"　",如发生|发展|理　。

4. **信息检索**(Information Retrieval,**简称**IR)**和信息摘录**(Information Extraction,**简称**IE)　信息检索是根据用户的查询要求从存有多达数百万份文档的文本数据库中搜索出相关的文档来。由于文本中的实词是信息的有效载体,因此以词或短语作为文档的标引项或检索项更合理。在印欧语文本中,词间有空格作为分隔标记,采用关键词来实现标引和检索在技术上不存在任何障碍。对汉语来说,情况就不同了。大多数汉语信息检索系统是以词为基础的,但也有个别系统为了回避大规模真实文本的分词困难,不得不采用以汉字(或汉字串)为基础的标引与检索方法。这种方法虽然可以保证较高的查全率,但缺点是查准率低,检索速度慢,而且无法利用同义词或上义词来扩充搜索范围等。比如,假设检索项是"华人",机器会误把含有"中华人民共和国"字串的文档也列入相关文档,从而导致查准率的下降。

信息摘录要求计算机从自然语言输入文本中摘出事先指定的信息项,从而自动生成某个领域的数据库。比如,根据一篇自然灾害的新闻报道,可摘录出灾害的类型(地震、洪水或森林火灾等),灾害发生的时间、地点,人员伤亡数字和救援情况等。这类信息处理任务也必须以词或短语作为处理对象,因此也离不开自动分词技术。

5. **文本分类和自动文摘**　文本分类和自动文摘在政府部门和企事业单位的办公自动化和文档管理中应用日趋广泛,尤其是当互联网(Internet)已在全球大规模普及的情况下,无论从防范有害信息入侵还是从有效利用有益信息的角度来看,对网上广为传播的形形色色的文本实行分类与过滤都是紧迫和必要的。

文本分类和自动文摘都需要抽取出每类文本和每个文本的某些特征来,这些特征一般可以以文本中有区别意义的字或词(语)来体现。如果输入文本的特征和某一类文本的特征最大程度地匹配,机器就确认它属于这类文本。如果输入文本的某个句子或段落集中出现这样的特征字或词(语),就把它们摘录下来作为这个文本的摘要。由于处理中依据的仅仅是文本最

表层的字 - 词信息,还谈不上对文本的理解,所以人们把它们分别叫做机械分类和机械文摘。从发展的眼光来看,只要汉语自动分词技术能在精度和速度上满足网上大规模文档分词任务的要求,那么无疑以词(语)为基础的文本分类和自动文摘技术将优于以字为基础的技术。因为只有以词(语)作为文本的特征,词性、词义和句法结构等更深层的语言知识才有用武之地。换句话讲,要想从当前的机械分类和机械文摘过渡到基于理解的文本分类和自动文摘,只有基于词(语)的技术才具备这种可能性。

6. **汉字的侦错与纠错**　目前在国内销售和使用的计算机都配备有中英文的文本编辑器(editor)。英文编辑器大都带有一个拼写检查程序(spelling checker),帮助用户检查英文文本中出现的拼写错误。这种程序依靠一部在线的英文词典和一些词法规则便可以完成拼写错误的检查。这种工作方式对汉语文本并不适用,因为不论用何种输入方式(如键盘输入、汉字识别输入或语音输入等),显示在计算机屏幕上的每个汉字都必然是汉字编码字符集中的一个单字,它可能是别字或冗余字,但不会是缺一点或少一撇的错字。因此,要实现汉字的侦错与纠错,也要求对文本进行自动分词。如果一个字在某些词(语)中的出现是不合理的,机器就判定它是别字。如"一鸣惊人"中的"鸣"是别字,应更正为"鸣";"文木"中的"木"字可能是"本"字等等。校对系统一般都要先对输入文本实行自动分词,然后凭借大规模语料库的字 - 词同现统计信息来实现侦错 - 纠错功能。

上述技术还大量应用于汉字识别或语音识别系统的后处理中,以提高识别精度。在这些识别系统中,一般都会在识别过程中对当前待识别的汉字给出若干候选字,系统利用类似汉字侦错 - 纠错的技术在这些候选字中自动选定一个最恰当的字。据报道,具有上述后处理能力的识别系统大体上可以把识别精度提高 10 %左右。

7. **词语的计量分析**　词语计量分析已广泛应用于词频统计、新词调查、计算机辅助词典编纂、词语搭配研究和文章或作者的风格学(stylistics)研究等众多领域。这方面的成果有 1985 年北京语言学院在人工分词的基础上对拥有 131 万词次的现代汉语语料所进行的词频统计,统计结果公布在《现代汉语频率词典》中。[4]又如,被国际辞书出版界誉为全世界第一部用计算机编纂的大型辞书——《COBUILD 英语词典》,[5]是在英国伯明翰大学和考林斯出版社合作建造的总容量为 2000 万词次的 COBUILD 语料库支持下完成的。不仅入选该词典的每个词条都以 COBUILD 语料库的词频数据为依据,而且词条的每一种用法(或义项)及其相应例句都拥有语料库提供的实证。一部辞书的例句不是由编者生造的,而是利用计算机从大规模语料库中选取的,单是这一条就足以说明 COBUILD 词典的卓越贡献了。

8. **自然语言理解**　语言信息处理的最高目标大概就是自然语言理解了。自然语言人 - 机接口,问答系统,基于理解的文本分类、自动文摘、信息摘录和自动翻译等,都是自然语言理解的极具应用价值的实例。在这些众多的应用领域中,对输入文本进行句法分析(parsing)恐怕是必不可少的一项处理任务。由于词是"最小的能独立运用的语言单位",句法分析的前提也是自动分词。因为计算机从事句法分析所凭借的语言知识不外乎来自机器词典和句法规则。词典中收录了每个词条的词法、句法和语义知识,具体包括词条的词性,词义解释或语义分类,动词的论元结构(argument structure,又称配价)、论旨属性(thematic property,即格框架及其中每个论旨角色的选择性语义限制)和适用句型等等。而句法规则一般来讲是在词类等知识的基础上编制的。因此,如果输入语句未经分词处理,就无从根据句中出现的每个具体词

到机器词典中去查找相应的语言知识;而如果不知道每个具体词的词性等词汇知识也就不可能直接调用相关的句法规则来正确判断短语或句子的句法结构。

综上所述,在汉语信息处理众多的应用领域中,从最底层的各类汉字输入方法到最高层的汉语理解研究,都不可能完全回避汉语文本的分词问题。所以在当前把分词问题看作制约汉语信息处理发展的一个瓶颈并非言过其实。说它是瓶颈,不仅因为其重要,而且因为其困难。

二 汉语分词的困难

汉语分词所面临的困难可以从如下四方面来考察:

1. "词"这个概念在汉语中是否有清晰的界定? 大多数以汉语为母语的人对于何谓词充其量只有一种朦胧的感觉。换句话讲,词这个概念对他们来说缺乏一种心理实在性(psychological reality)。在汉语中,词对下同语素之间、对上同短语(即词组)之间往往没有清晰的界限。有关分词标准的问题,给出一条定义或颁布一个分词规范还不能从根本上解决问题。[6]比如,在《现代汉语词典》(修订本)(以后简称《现汉》)中,"养"的第5个义项被释义为"培养"。问题是:在这个义项下,"养"究竟是一个词还是一个语素呢?说它是词,似乎又不能单说;说它不是词吧,又解释不了为什么这部词典不收出现频度颇高的合成词"养成"。又如"酿"的第3个义项为"逐渐形成",据笔者判断这个"酿"是一个语素而不是词,据此推理,频度略低于"养成"的"酿成"应当是一个词而不是动补短语,为什么《现汉》的编者也不将它作为合成词收入词典呢?类似这样的例子还很多,这至少说明即使我国语言学界对于词的概念也有点飘忽不定。我们迄今拿不出一个公认的、具有权威性的词表来,在这种情况下让计算机去实现自动分词岂不是有点勉为其难了吗?这是分词问题所面临的第一个困难。

2. 分词和理解,孰先孰后? 即使汉语有了一个公认的词表,计算机分词仍然面临知识短缺的大问题。从我们自身的体验来看,人在阅读(或朗读)一篇汉语的文章时,大体上是先理解后分词,或至少是边理解边分词。断词断不下来而需要返回去重新读的情况也偶尔有之。举例说,像:

美国会通过赫尔姆斯—伯顿法。(1)

发展中国家兔的计划。(2)

这样的句子,难道人读时就没有磕磕碰碰的现象吗?然而计算机大概永远做不到像人那样先理解后分词,而只能是先分词后理解,因为计算机理解文本的前提是分词在先(详见上一节"自然语言理解")。由此看来,计算机自动分词问题颇有点像究竟是鸡先生蛋还是蛋先生鸡这样的怪圈,即分词要以理解为前提,而理解又要以分词为前提。如果我们承认分词系统只能在对输入文本尚无理解的条件下进行分词,那么我们不仅只能在相当有限的表层知识支持下来考虑分词算法,而且决不可企求百分之百的正确切分。这是自动分词所面临的第二个困难。

3. 歧义切分字段 自动分词所面临的第三个困难是文本中的歧义切分字段。歧义字段的情况相当复杂,仅举两种典型情况,即交集型和包孕型(又称组合型或多义型)歧义字段。

(1)交集型歧义字段 假设A,B,C分别代表由一个或多个字组成的字串,如果在ABC字段中A,AB,BC,C分别都是词表中的词,则称该字段为交集型歧义字段。显然如果仅根据词表的知识,那么AB/C和A/BC都是合理的切分结果。如果不向分词系统提供进一步的句法-语义知识,系统很难从这两种切分结果中作出正确的抉择。以上例句(1)中的"美国会"字

段是交集型的,它可产生"美/国会"和"美国/会"两种切分结果。例句(2)属于更复杂的歧义字段,参见[7]。据统计,交集型歧义字段占全部歧义字段的85%以上,[8]是自动分词系统需要重点加以解决的疑难问题,因为它对分词的正确率有很大影响。

(2)包孕型歧义字段　在字段AB中,如果A,B,AB分别都是词表中的词,则称AB为包孕型歧义字段。例如:

明天/她/将/来/北京/。(3a)

税收/制度/将来/会/更/完善/。(3b)

依靠/群众/才/能/做/好/工作/。(4a)

现在/是/施展/才能/的/好/机会/。(4b)

把/手/举/起来/。(5a)

茶杯/的/把手/断/了。(5b)

此外,像"十分""个人""马上""学会""了解"等也都是包孕型歧义字段。

4. 未登录词的辨识　这是自动分词所面临的第四个困难。许多分词算法都是在完备词表的假设下设计的,其实这一假设并不成立。汉语和其他自然语言一样,它的实词部分永远是一个开放集,不但因为社会上的新词将不断涌现,而且专有名词虽然不新,但不可能尽收,如人名、地名、机构名、译名等等。未登录词造成的分词错误远远超过歧义切分字段引发的错误,因此近年来这个问题已成为自动分词研究的焦点。[9]~[11]下面的例句选自[12]:

(正) 王建国/执勤/的/岗位/是/地处/闹市/的/解放路/。(6a)

(误) 王/建国/执勤/的/岗位/是/地处/闹市/的/解放/路/。(6b)

(正) 随同/穆巴拉克/总统/来访/的/有/副总理/阿斯马特·阿卜杜勒·马吉德/…(7a)

(误) 随同/穆/巴/拉/克/总统/来访/的/有/副总理/阿/斯/马/特/·/阿/卜/杜/勒/·/马/吉/德/…(7b)

(正) 任/何庆宝/怎么/讲/,/他/也/不/信/。(8a)

(误) 任何/庆/宝/怎么/讲/,/他/也/不/信/。(8b)

以上讲了汉语分词的四个困难。其中,第一条是分词标准问题,是人和计算机共同面临的困难。如果在分词标准上不能尽快达成共识,那么在词表和带标注的语料库等重要资源上就不能做到共享与复用,势将造成重复开发。第二条阐明了一个严酷的事实,说明计算机不可能像人那样先理解后分词,因此自动分词只能是在知识严重短缺的条件下去追求比较好的分词结果。第三和第四条是影响分词精度的两个主要因素,因而也是当前自动分词研究的焦点。

三　一种最基本的分词方法——MM法

要想大致了解一下计算机是怎样分词的,最好先从最大匹配法(Maximum Matching Method,简称MM法)入手。有兴趣了解更多分词方法的读者可参阅[13][14]。

1. 最大匹配法　MM法是一种得到广泛应用的机械分词方法,说它"机械",因为它在分词过程中除了依靠一个分词词表以外不再拥有其他词法、句法和语义知识。MM法的基本过程是这样的:假设词表中最长的词由i个字组成,则每次从句子头上截取一个长度为i的字串,令它同词表中的词条依次匹配,如果词表中确有这样一个i字词,匹配成功,就把这个字串

作为一个词从句子头上切分出去。然后再从句子余下部分的头上截取另一个 i 字字串,重复上述过程,直至句子被切分完为止。如果在词表中找不到一个词条能同当前字串匹配,就从该字串尾部删去一个字,用 i - 1 字长的字串到词表中去查找,若匹配成功同样把该字串作为一个词从句子中切分出去;若匹配失败,从该字串尾部删去一个字,再用 i - 2 的字串去词表中匹配,直至匹配成功。例如输入句子是"中华人民共和国成立了"。假设词表中有"中华人民共和国"这个词,而且词表中最长的词也是 7 个字(即 i = 7),则第一次从 句首截取的 7 字字串"中华人民共和国"就匹配成功。句子余下部分为 3 字字串"成立了",词表中没有这样的 3 字词,字串截尾得新字串"成立",匹配成功。句子余下 1 字串"了",也匹配成功。于是句子被切分为"中华人民共和国/ 成立/ 了"。注意,虽然词表中也有"中华、人民、共和、国"等多字词和"中、人、共、和、国"等单字词,MM 法不会把字串"中华人民共和国"中的短词切分出来。然而 MM 法的这个特点也会在长词覆盖短词的情况下引起错误切分。例如,当输入句子是"有个人叫张梦云",MM 法将把"个人"作为一个词切分出来,得到"有/ 个人/ 叫/ 张梦云"这样的错误切分。这是因为 MM 在给出一套可以重复运作的切分流程的同时,也包含了"长词优先"的切分评估原则。即认为对同一个句子来说,切分所得的词数最少时是最佳切分结果。虽然这一评估原则在大多数情况下是合理的,但也会因此引发一些切分错误,详见 2。

MM 法根据对当前字串的扫描方向又分为正向最大匹配(FMM)法和反向最大匹配(BMM)法。以上讲的是 FMM 法。如果从当前字串的尾开始从右到左扫描,便是 BMM 法。这时,若匹配失败,要从候选字串中去掉最前头的一个字以形成新的候选字串。

2. 利用 FMM 和 BMM 法检查歧义切分

利用 FMM 和 BMM 法切分同一文本,可以起到核对切分结果的作用。据孙茂松、邹嘉彦报道,[15]对新闻语料中随机选出的 3680 个句子进行测试,得出以下四种情况的统计数据·

情况(1) FMM 和 BMM 的切分结果不同,但两种结果均不正确,只有 2 句,占测试句子总数的 0.054 %·

以新的姿态出现在世界东方 (9)

(误) 以/ 新/ 的/ 姿态/ 出现/ 在世/ 界/ 东方 (FMM)

(误) 以/ 新/ 的/ 姿态/ 出/ 现在/ 世界/ 东方 (BMM)

情况(2) FMM 和 BMM 的切分结果不同,其中有一种结果正确,有 340 句,占测试句子总数的 9.24 %:

使节约粮食进一步形成风气 (10)

(误) 使节/ 约/ 粮食/ 进一步/ 形成/ 风气 (FMM)

(正) 使/ 节约/ 粮食/ 进一步/ 形成/ 风气 (BMM)

情况(3) FMM 和 BMM 的切分结果相同,但不正确,有 15 句,占测试句子总数的 0.41 %·

反映了一个人的精神面貌 (11)

(误) 反映/ 了/ 一/ 个人/ 的/ 精神/ 面貌 (FMM,BMM)

情况(4) FMM 和 BMM 的切分结果相同,而且正确,共有 3323 句,占测试句子总数的 90 30 %:

美国加州大学的科学家发现 (12)

(正) 美国/ 加州/ 大学/ 的/ 科学家/ 发现 · (FMM ,BMM)

根据上述实验可得出如下结论:

1) 对比 FMM 和 BMM 的切分结果可以作为一种简易可行的方法来判断切分是否正确。从情况(3)和(4)可知,当两种方法的切分结果相同时,有大约 99 %的把握认为切分结果是正确的。而从情况(1)和(2)可知,当两种方法的切分结果不同时,也有大约 99 %的把握认为其中至少有一种切分结果是正确的。当然要想判明哪一种结果正确,还有赖于其他语言知识。

2) 虽然上述情况(1)和(3)的出现概率不到 0.5 %,但它说明 MM 法存在着切分的盲区,即不管 FMM 和 BMM 的切分结果相同还是不同,它们可能都没有找到正确的切分。这正是 MM 法的一个弱点。

四 结 语

汉语信息处理是一项庞大的系统工程,其中涵盖了从字、词、短语、句子、语篇等多层面的信息加工处理任务。80 年代随着计算机汉字输入输出问题的基本解决,汉语信息处理的主战场已从"字处理"转移到"词处理"。尽管从全局来看,我们还有很长的一段路程要走,但是即使在"词处理阶段"也可以找到许多有广阔前景的的应用领域。当前迫切需要解决的一个问题就是汉语文本的自动分词,在这个问题中又必须首先对汉语的分词标准取得共识。分词规范作为一项国家标准已公布 4 年了,除了个别条款需要修订以外,现在的当务之急是制定一部与之相配合的词表。相信用规范加词表的方式可以把汉语的分词标准明确下来。

本文的用意是阐明分词在汉语信息处理中的重要作用及其面临的困难,希望语言学界和信息界有更多的同事来关心和参与这个问题的解决。

参考文献

[1]《汉语信息处理词汇 01 部分 :基本术语(GB12200.1 - 90)》,中国标准出版社,1991。

[2]高家莺、范可育、费锦昌《现代汉字学》,高等教育出版社,1993。

[3]王永成等《中文信息处理技术基础》,上海交通大学出版社,1991。

[4]王还、常宝儒等《现代汉语频率词典》,北京语言学院出版社,1986。

[5]Sinclair J Collins COBUILD English Language Dictionary ,London and Glasgow :Collins ,1987

[6]《信息处理用现代汉语分词规范(GB/ T13715 - 92)》,中国标准出版社,1992。

[7]侯敏、孙建军《汉语自动分词中的歧义问题》,《语言文字应用》1996 年第 1 期(总第 17 期)。

[8]梁南元《书面汉语自动分词系统 —CDWS》,《中文信息学报》1(2) ,1987。

[9]张俊盛等《多语料库作法之中文姓名辨识》,《中文信息学报》6(3) ,1992。

[10]孙茂松等《中文姓名的自动辨识》,《中文信息学报》9(2) ,1995。

[11]孙茂松、张维杰《英语姓名译名的自动辨识》,陈力为主编《计算语言学研究与应用》,北京语言学院出版社,1993。

[12]Sun Maosong ,Huang Changning. Word Segmentation and Part of Speech Tagging for Unrestricted Chinese Texts :A Tutorial. ICCC'96 ,Singapore ,June 4 ,1996

[13] 揭春雨、刘源、梁南元《汉语自动分词方法》,《中文信息学报》3(1) ,1989。

[14] 陈群秀《国内汉语自动分词研究进展》,《计算机世界报》1992 年 4 月 22 日(总第 387 期)第 77 版。

[15] Sun Maosong ,Benjamin K. T'sou. Ambiguity Resolution in Chinese Word Segmentation ,Benjamin K. T'sou & Tom B Y Lai (eds.) Proceedings of PACLIC - 10 ,Hong Kong ,27 - 28 Dec. , 1995

(黄昌宁 清华大学计算机科学与技术系,邮编:100084)